DOCUMENT RESUME

ED 401 323 TM 025 903

AUTHOR Wang, Tianyou; And Others
TITLE Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT.
PUB DATE Apr 96
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Ability; *Classification; *Error of Measurement; Goodness of Fit; *Item Response Theory; Performance Factors; *Probability; Reliability; *Scaling; Statistical Distributions; *True Scores
IDENTIFIERS Polytomous Variables; Work Keys (ACT)

ABSTRACT

M. J. Kolen, B. A. Hanson, and R. L. Brennan (1992) presented a procedure for assessing the conditional standard error of measurement (CSEM) of scale scores using a strong true-score model. They also investigated the ways of using nonlinear transformation from number-correct raw score to scale score to equalize the conditional standard error along the reported score scale. Kolen, L. Zeng, and Hanson (in press) presented a similar procedure for assessing CSEM using item response theory (IRT) techniques. This paper extends that procedure to tests with polytomous items using a polytomous IRT model approach. A polytomous IRT-based procedure to assessing decision consistency of performance level classification based on alternate test forms is also described. The general approach for assessing CSEM and reliability is to obtain the probability distribution of the level score conditioned on a given theta and then compute the conditional mean and conditional standard deviation (or variance) of the scale scores or the level scores. The CSEM of the level score is the conditional standard deviation. Data from the American College Testing Program's Work Keys assessment for 7,097, 2,035, and 1,793 samples illustrate the use of the procedures. Model fit, classification consistency, and reliability were evaluated and found acceptable. Results suggest that the new level scores have higher reliability and classification consistency than the old level scores, indicating the usefulness of these polytomous IRT-based procedures. (Contains 4 tables, 5 figures, and 16 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 323

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
TIANYOU WANG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT

Tianyou Wang, Michael J. Kolen, Deborah J. Harris

Measurement Research Department

ACT

A paper presented at the 1996 Annual Meeting of the American Educational Research Association.

Address correspondence to Tianyou Wang, P.O.Box 168, Iowa City, IA 52243.

TM025903

BEST COPY AVAILABLE

Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT

Tianyou Wang, Michael J. Kolen, Deborah J. Harris

ACT

Performance assessments items are usually scored on a polytomous score scale. In some testing programs (e.g., Work Keys, ACT 1995), the final reported scores are on a proficiency level type of scale, i.e., the examinees are classified into a finite number of levels of proficiency. Classifications are often based on converting raw scores to levels, because it is relatively easy to use. In other testing programs, raw scores are converted to a reported score scale using some linear or non-linear transformations. In either case, it is useful to obtain and report information about the conditional standard error of measurement (CSEM, conditioned at each level score or scale score), and the overall reliability. In the case of performance levels, it is also helpful to report information about classification decision consistency. To provide test users with the above information is in accordance with the recommendation by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985), especially, Standards 2.10, 2.12, and 11.3.

Kolen, Hanson, and Brennan (1992) presented a procedure for assessing the CSEM of scale scores using a strong true-score model. In that article, they also investigated the ways of using non-linear transformation from number-correct raw score to scale score to equalize the conditional standard error along the reported score scale -- a property that facilitates score interpretation. Kolen, Zeng, and Hanson (in press) presented a similar procedure for assessing the CSEM, but using IRT techniques. Both these procedures were primarily developed for tests with dichotomously scored items. One purpose of this paper is to extend the procedure described in Kolen et al. (in press) to tests with polytomous items using a polytomous IRT model approach and to discuss the similarity and difference between scale scores and level scores.

Another main purpose of this paper is to describe a polytomous IRT-based procedure for assessing decision consistency of performance level classification based on alternate test forms.

Usually, assessing decision consistency requires a data collection design that would require each examinee to take more than one test form. With this type of data, for a test of L performance levels, a L X L contingency table can be constructed based on examinees' performance levels on two test forms. Subkoviak (1984) presented some decision consistency indices which can be computed based on the contingency table. In many testing situations, however, this type of data collection design is not feasible due to time constraints or other conditions. With each examinee only taking one test form, conventional methods for assessing decision consistency do not readily apply. IRT-based techniques, however, can be used to overcome this difficulty.

In the next section, we will describe two polytomous IRT-based procedures. The first procedure, which can be used assess CSEM and reliability, applies to both scale scores and performance level scores. The second procedure, which can be used to assess decision consistency, only applies to performance level scores. After the descriptions, some examples are given using the Work Keys data to demonstrate the usefulness of these procedures.

IRT Procedure for CSEM and Reliability

The general approach for assessing the CSEM and reliability is the same as the procedure described in Kolen et al. (in press). The central task is to first obtain the probability distribution of the level score (or scale score) conditioned on a given θ and then compute the conditional mean and conditional standard deviation (or variance) of the scale scores or the level scores. The CSEM of the level score is the conditional standard deviation. Given a θ distribution for an examinee population, conditional means and conditional variances can be integrated over the θ distribution to obtain the overall error variance and true score variance. Reliability can thus be computed based on this information. The main difference between the present procedure and the one described in Kolen et al. (in press) lies in the step for obtaining the conditional level score distribution. In the case of scale scores, usually the scale score are converted from the raw total score using some non-linear conversion table. In the case of level scores, the derivation of level scores may not be direct from raw total score. In the example of Work Keys tests, the conversion was originally based on a

ninth order statistic from the 12 ratings given by two raters on six items. In this paper, we will describe in detail the computation procedure for level scores derived from the raw total scores and will provide some general guidelines for computing conditional standard errors for level scores that are derived not from raw total scores.

The Polytomous IRT Probability Models

Various polytomous IRT models have been developed: nominal response model (Bock, 1972), rating scale model (Andrich, 1978), graded response model (Samejima, 1969), partial credit model (Masters, 1982), and generalized partial credit model (Muraki, 1992), etc. With any of these models fitted to the polytomous test data, the probability of getting a particular response on an polytomously scored items can be computed given a θ value. In the present paper, the (generalized) partial credit model is used to fit the test data, though the polytomous IRT-based procedures described in this papers apply with any of the models mentioned above. Let U_k be random variable for the score on item k with scores from 0 to m, with the generalized partial credit model, the probability of getting a particular response j is given by

$$\Pr(U_k = j|\theta) = \frac{\exp\left[\sum_{v=0}^{j} a_k(\theta - b_k + d_v)\right]}{\sum_{c=0}^{m}\exp\left[\sum_{v=0}^{c} a_k(\theta - b_k + d_v)\right]}, \qquad (1)$$

where a_k is the discrimination parameter, b_k is the difficulty parameter, and d_{kv} (v=0, 1, ... m) are the category parameters for item k.

Conditional Distribution of Raw Total Scores

Assume there are K polytomous items and let U_k be random variable for the score on item k $(U_k = 0,1,\cdots n_k)$. Let $\Pr(X = x|\theta)$ $(x = 0,1,\cdots,T)$ represents the conditional distribution of the raw total score. For dichotomous items, this distribution is a compound binomial distribution as indicated by Lord (1980). Lord and Wingersky (1984) provided a recursive algorithm for computing this distribution. For polytomous items, this distribution is a compound multinomial distribution. Hanson (1994) extended the Lord-Wingersky algorithm to the

polytomous items. (The same extension was also provided by Thissen, Pommerich, Billeaud, and Williams, 1995). This recursive algorithm is described as the following:

For item 1,

$$\Pr_1(X=x|\theta)=\Pr(U_1=x|\theta),\quad x=0,1,\cdots,n_1. \tag{2}$$

For item $k=2,\cdots,K$,

$$\Pr_k(X=x|\theta)=\sum_{u=0}^{n_k}\Pr_{k-1}(X=x-u|\theta)\Pr(U_k=u|\theta), \tag{3}$$

for $x=0,1,\cdots,\sum_{j=1}^{k}n_j$. $\Pr(U_k=u|\theta)$ is given by equation 1 if a generalized partial credit model is used. The total raw score distribution is obtained after all the K items are included.

With this algorithm, we can compute the conditional distribution $\Pr(X=x|\theta)$ $(x=0,1,\cdots,T)$, where $T=\sum_{k=1}^{K}n_k$.

Conditional Distribution of Level Scores

Let S symbolizes the raw-to-level transformation, following the same logic as in Kolen et al. (in press), the conditional distribution of the level score can be expressed as:

$$\Pr(S(x)=s|\theta)=\sum_{x:S(x)=s}\Pr(X=x|\theta),\quad s=1,2,\cdots,L \tag{4}$$

The mean and variance of the conditional level score distribution are:

$$\xi(\theta)=E[S(X)|\theta]=\sum_{s=1}^{L}s\Pr(S(x)=s|\theta) \tag{5}$$

$$\sigma^2[S(X)|\theta]=E\left\{[S(X)-\xi(\theta)]^2|\theta\right\}=\sum_{s=1}^{L}s^2\Pr(S(x)=s|\theta)-\xi(\theta)^2 \tag{6}$$

The conditional mean is just the conditional true level score; the square root of the conditional variance is the CSEM conditioned on θ scale. To find the CSEM conditioned on the

level score, let $s = \xi(\theta)$ and express θ in terms of level score: $\theta = \xi^{-1}(s)$. Substituting this expression in equation 6 yields the CSEM conditioned on level score s.

If the level scores are not converted directly from the raw total scores, then the conditional distribution of the level scores must be obtained by the other approaches. If the theoretical distribution of the statistic that is used to compute the level scores is known, then a theoretical approach can be used. In settings where theoretical approaches are not available, simulation techniques can be used to estimate the conditional distribution of the level scores. The steps after that are the same as those described by equations 5 and 6. The simulation technique to obtain the conditional level score distribution is illustrated in a later part of this paper.

Average Error Variance and Reliability

Following the same logic used by Kolen et al. (1992), and Kolen et al. (in press), average error variance and reliability of the total raw scores and level scores can be computed if a θ distribution is given. To summarize their procedure, the following equations are presented here.

Let *E* denote error scores, $\Psi(\theta)$ denote the distribution θ, and the subscript *S* denote level scores. The average error variance is given by

$$\sigma^2(E) = \int_{\theta} \sigma^2(X|\theta)\Psi(\theta)d\theta \qquad (7)$$

The marginal distribution of the raw total score is obtained by

$$\Pr(X = x) = \int_{\theta} \Pr(X = x|\theta)\Psi(\theta)d\theta \qquad (8)$$

The observed raw score variance $\sigma^2(X)$ is the variance of this marginal score distribution. The reliability of raw total scores is given by

$$rel_{raw} = 1 - \frac{\sigma^2(E)}{\sigma^2(X)} \qquad (9)$$

Similarly, average error variance for the level scores and observed score variance for level scores can be computed also by substituting *S* for *X* in equations 7 and 8.

$$\sigma^2(E_s) = \int_{\theta} \sigma^2(S|\theta)\Psi(\theta)d\theta \tag{10}$$

$$\Pr(S=s) = \int_{\theta} \Pr(S=s|\theta)\Psi(\theta)d\theta \tag{11}$$

The observed level score variance $\sigma^2(S)$ is the variance of this marginal score distribution. The reliability of the level scores is then given by

$$rel_{scale} = 1 - \frac{\sigma^2(E_s)}{\sigma^2(S)} \tag{12}$$

IRT Procedure for Decision Consistency

Decision consistency is an important type of reliability concept for educational assessment that involve classification decisions. Previous literature in this area mostly dealt with dichotomous classifications, namely mastery and non-mastery. But most of the procedures and indices developed for dichotomous classifications can be easily extended to classifications with polytomous levels. In the next part, we will first discuss some important decision consistency indices and some procedures for calculating these indices and then we will describe in detail an IRT-based procedure to assess decision consistency for polytomous classifications.

Decision Consistency Indices

Subkoviak (1984) reviewed some decision consistency indices and procedures to compute them. Based on his review, two most important indices are p_0 and κ. p_0 is the proportion of consistent classification based on two parallel forms and κ is the proportion of consistent classification adjusted for chance. These two indices are computed based on a classification outcome table (we will refer to it as the contingency table) as illustrated by the following hypothetical example of 3-level classification:

		Form 2			
		Level 1	Level 2	Level 3	Low total
Form 1	Level 1	0.12	0.11	0.07	0.30
	Level 2	0.09	0.23	0.08	0.40
	Level 3	0.03	0.10	0.17	0.30
	Column Total	0.24	0.44	0.32	1.00

With this table, consistent proportion $p_0 = \sum_{k=1}^{L} \hat{p}_{kk} = 0.12 + 0.23 + 0.17 = 0.52$. In order to compute κ, we must first compute p_c which is the consistent proportion due to pure chance, that is, if the two forms are independent and the entries of the table would be the product of the corresponding row and column totals. So $p_c = \sum_{k=1}^{L} \hat{p}_{k.} \hat{p}_{.k} = 0.30*0.24 + 0.40*0.44 + 0.30*0.32 = 0.346$. And

$$\kappa = \frac{p_0 - p_c}{1 - p_c} = \frac{0.52 - 0.346}{1 - 0.346} = 0.266 \ . \qquad (13)$$

Subkoviak (1984) reviewed several procedures for obtaining the contingency table. The simplest and most direct way is to administer two parallel forms to the same group and then summarize the outcome of classification based on the two test scores into the contingency table. If a two-administration data collection design is not feasible, then some statistical model-based approximation procedures can be used as substitutes. For examples, Huynh (1976) proposed a complicated method based on Keats and Lord's (1962) beta-binomial model. Subkoviak (1976) proposed a simpler but less accurate method which uses binomial distribution for approximation. Hanson and Brennan (1990) proposed a method using the strong true score model. These methods might be extended to polytomous classification case. In the next part, we will present an IRT-based procedure to fulfill this task with reasonable accuracy and manageable computation provided that IRT assumptions are met for the test.

IRT Procedure for Obtaining the Contingency Table

Earlier in this paper, we have presented procedures for computing the conditional distribution of the level scores based on the item parameters and the classification criterion. Here it is assumed that the conditional distribution of the level scores are known. By the assumption of local independence, the responses to items in two parallel forms are independent conditioned on a single θ value, thus the classifications based on the two forms are also independent given that θ. Based on this result, a conditional contingency table can be computed by multiplying the corresponding conditional level score probabilities. Let $\left[A_{ij}|\theta\right]$ denote the entry *i, j* of the conditional contingency table, it is expressed as the following:

$$\left[A_{ij}|\theta\right] = \Pr(S_1 = i|\theta)\Pr(S_2 = j|\theta)\ ,\ i = 1,2,\cdots,L,\ j = 1,2,\cdots,L\ . \qquad (14)$$

Given a θ population distribution $\Psi(\theta)$, the overall contingency table can be computed by integrating the conditional contingency table over $\Psi(\theta)$ using numerical integration.

$$\left[A_{ij}\right] = \int_{\theta}\left[A_{ij}|\theta\right]\Psi(\theta)d\theta\ ,\ i = 1,2,\cdots,L,\ j = 1,2,\cdots,L\ . \qquad (15)$$

After the contingency table is obtained, decision consistency indices p_0 and κ can be computed using the procedure illustrated in hypothetical example in the previous section.

Examples with Real Test Data

Tests and Test Scores

In this paper, the Work Keys (ACT, 1995) test data are used to demonstrate the computations and usefulness of the previously described IRT-based procedures. The Work Keys Listening and Writing tests each consist of 6 items. The response to each item is rated on a 0 to 5 rating scale by two independent raters. When the two ratings differ by more than one point, a third "expert" rater is used. The third rater's ratings replaces the ratings of each of the first two raters. Each examinee has 12 ratings. The original proficiency levels correspond to the ninth order statistic of the 12 ratings. That is, the 12 ratings are sorted from the highest to the lowest, and the ninth

highest score is the level score assigned to this examinee. This process was followed so that 75% of the 12 ratings earned by the examinee would be at or below the reported level. Because of the complexity of this classification procedure, classical test theory is not readily applicable in analyzing the psychometric properties of the level scores. Also, in the scale development process for the Work Keys tests, some alternative procedures for assigning level scores have been considered, one of which is to assign level score according to the rounded mean of the 12 ratings. It is also desirable to compare the psychometric properties of these different level score assignment procedures.

The analyses reported here used the Tennessee 1994 data for three forms (10, 11, and 12) Work Keys Listening and Writing tests. The sample sizes for these three forms are 7097, 2035, and 1793, respectively.

First, a partial credit polytomous model (Masters, 1982) was fit to each of the data sets using the FACETS (Linacre, 1989). Each item originally has 6 score categories (0 to 5). The sum of two ratings results in 11 score categories for each of the six items. Because other calibration programs such as PARSCALE (Muraki & Bock, 1993) and MULTILOG (Thissen, 1991) can not handle items with more than 10 categories, FACETS was the only option to use. Because the sample size for form 10cc is too large for FACETS, only half of the data set (every other examinee) was used in calibration. Notice that FACETS can only fit a partial credit model which is a Rasch type IRT model whereas PARSCALE and MULTILOG can fit a generalized partial credit model or a graded response model. But the IRT-based procedures described in this paper also apply to these two other types of models.

Because it is cumbersome to derive the conditional distribution of the ninth order statistic based on some probability model, a simulation technique was used to find the conditional distribution of the old (ninth order statistics-based) level score. The steps after that are the same as those for the new level score.

Computational Steps

For the new level scores, the computation follows these steps:

(1) Conditioned on a quadrature point on the θ scale, (a) the raw score distributions were estimated using the extended Lord-Wingersky algorithm in equations 2 and 3, (b) the level score distributions were estimated using equation 4 with the raw score-to-level score conversion table, (c) the conditional expected (true) level score and error variance were computed using equation 5 and 6, (d) the conditional 6x6 contingency tables were computed using equation 14.

(2) Using an empirical θ distribution based on the θ estimates from the FACETS output, the following overall indices were computed using numerical integration: (a) error level score variance, observed level score variance and true level score variance, and reliability (equations 10, 11, and 12), (b) the overall contingency table (equation 15) and classification indices (p_0 and κ) (equation 13), (c)the marginal distribution of the level scores (equation 11).

For the old level scores, the computation follows these steps:

1) Conditioned on a quadrature point on the θ scale, simulate responses to each of the six items for 200 simulees with the same theta. Each response, which ranges from 0 to 10, were broken into two ratings which range from 0 to 5 based on the rule that the two ratings can not differ more than one point. The ninth order statistic was used as the old level score. Based on these simulated data, (a) the level score distribution, (b) the conditional mean (true) level score and error variance (equations 5 and 6), (c) the conditional 6x6 contingency table (equation 14) were computed.

(2) Using an empirical θ distribution based on the θ estimates from the FACETS output, the following overall indices were computed using numerical integration: (a) error level score variance, observed level score variance and true level score variance, and reliability (equations 10, 11, and 12), (b) the overall contingency table (equation 15) and classification indices (p_0 and κ) (equation 13), (c)the marginal distribution of the level scores (equation 11).

Results and Conclusions

Model fit was partially checked by comparing the expected score distribution based on the model and the actual score distribution based on the test data. Figures 1 and 2 plot the fitted and observed total score distributions. It was found that the fitted distributions were close to the observed score distributions, suggesting that the model fits reasonably well.

Figure 3 contain plots of the conditional expected (true) level scores for the old and new levels. These plots consistently show that the new level scores are easier than the old level score, particularly at low levels. This result, which is confirmed by the marginal distributions of the old and new level scores contained in Table 1 and 2, is not surprising because the mean score corresponds to the 6th or 7th order statistic, which is easier than the 9th order statistic. Comparison between the estimated marginal level score distribution based on the IRT model and the observed marginal level score distributions was close, which suggest the model-based methods are quite accurate in estimating the level score distributions

The conditional standard errors (CSEM) of the old and new level scores are presented in Figures 4 and 5. Figure 4 plot CSEM along the θ scale whereas Figure 5 plot CSEM along the level score scale. These plots show that CSEM for the old and new levels have quite different patterns. The old level scores have big CSEM around level one. Generally, the old level scores have larger CSEM than the new level scores.

The classification consistency and reliability indices are summarized in Tables 3 to 4. Again, these results clearly suggests that the new level scores have higher reliability and classification consistency than the old level scores. This result is consistent with the findings for the CSEM. All these results show that the polytomous IRT-based procedures described in this paper are useful and reasonably accurate.

References

ACT (1995). Work Keys Assessments. Iowa City: ACT

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to the polytomous items.* Unpublished research note.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253-264.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for the mental test scores. *Psychometrika, 27,* 59-72

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (in press). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, ??,* ???-???.

Linacre, J. M. & Wright, B. D. (1993). *FACETS.* MESA Press: Chicago.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement,* 8, 452-461.

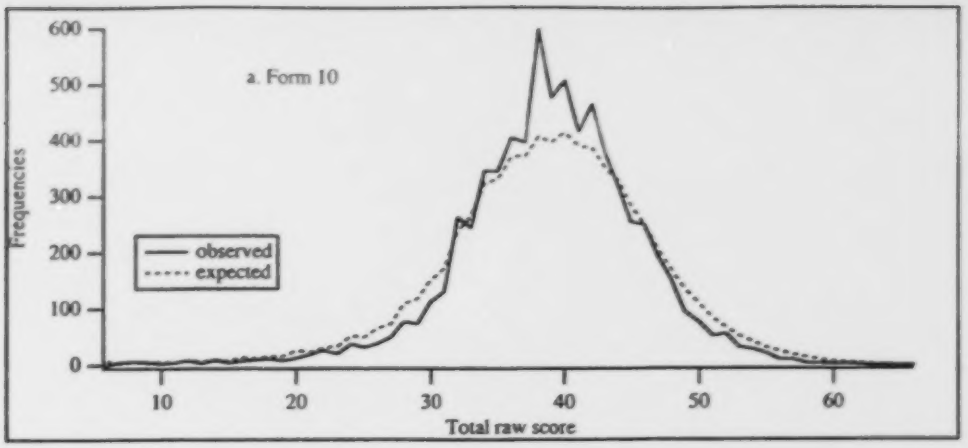
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

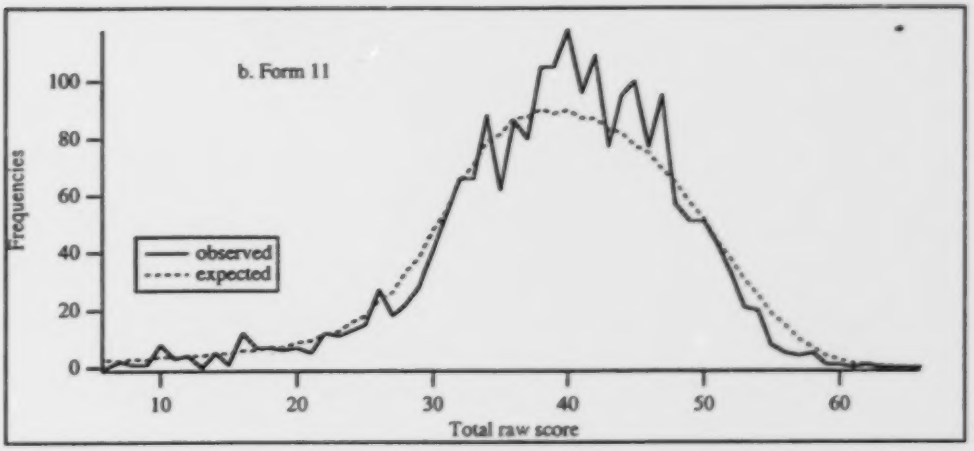
Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13,* 265-276.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In Berk, R. A. (Ed.). *A guide to criterion-referenced test construction.* Baltimore, MD: The John Hopkins University Press.

Thissen, D., Pommerich, M. Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.





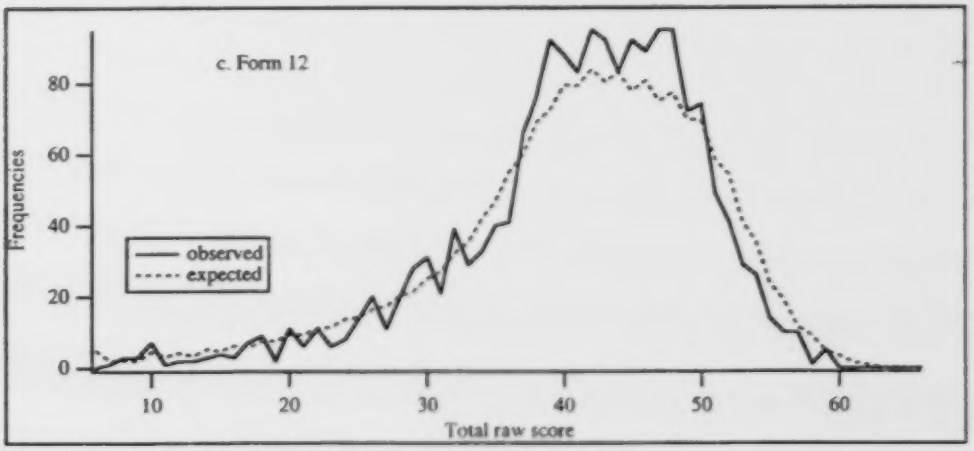
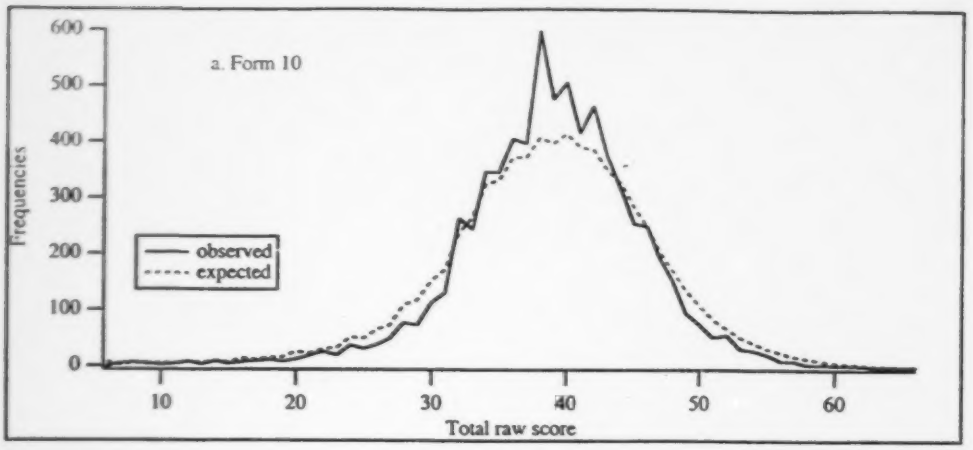
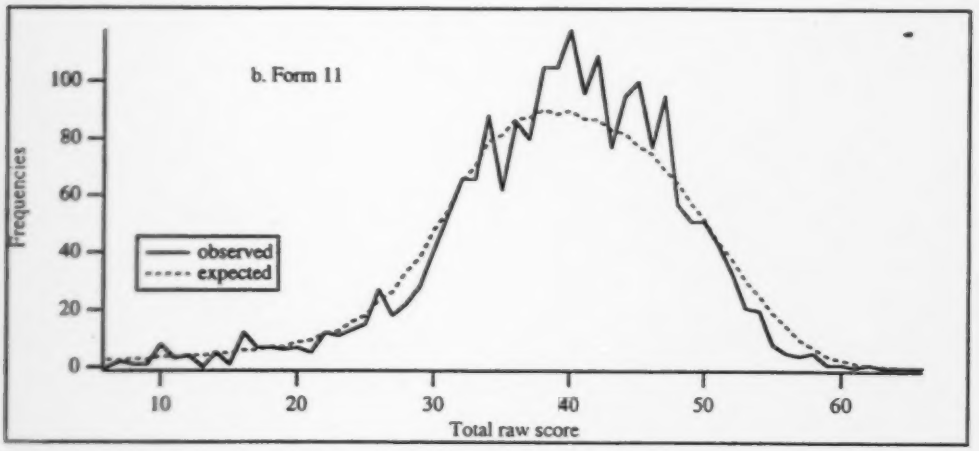


Figure 1. The fitted and observed score distribions for the three Listening forms.





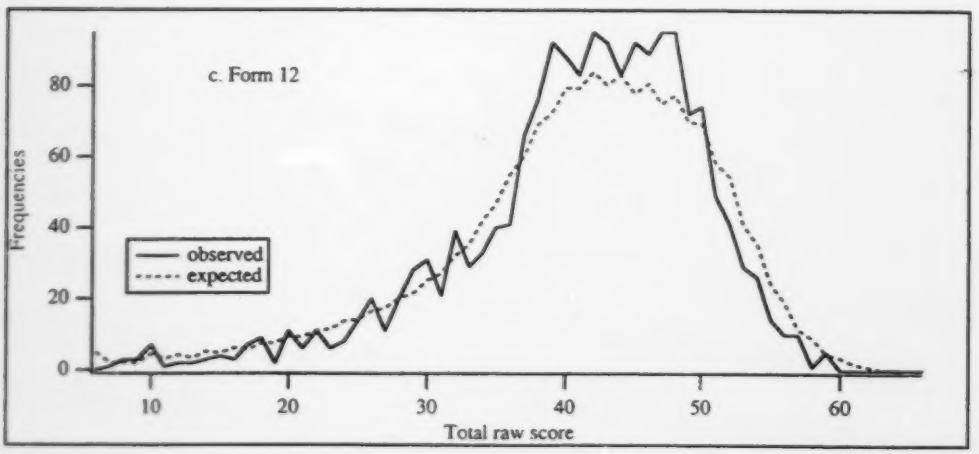


Figure 2. The expected and observed score distribions.

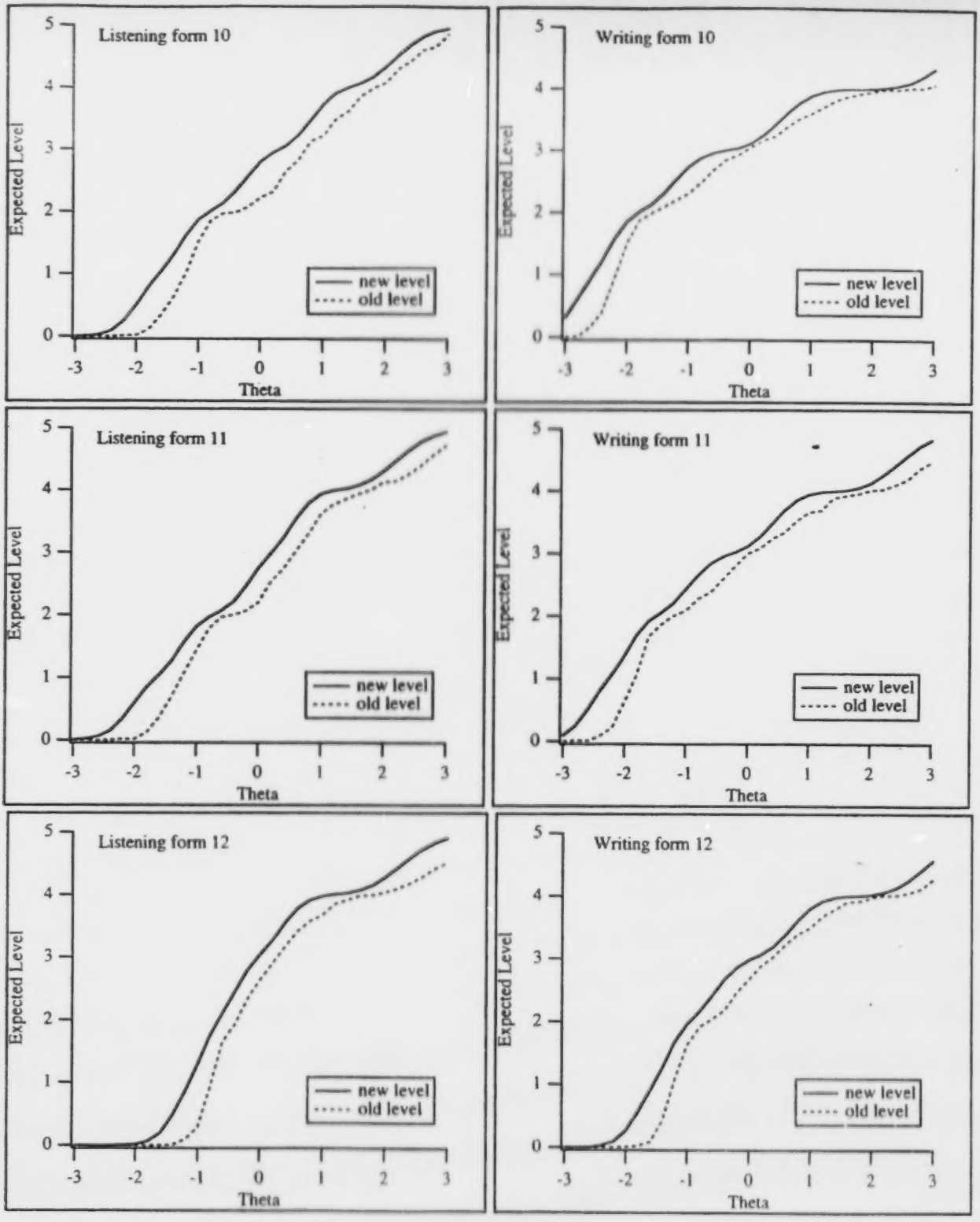


Figure 3. The conditional expected (true) level scores for old and new levels.

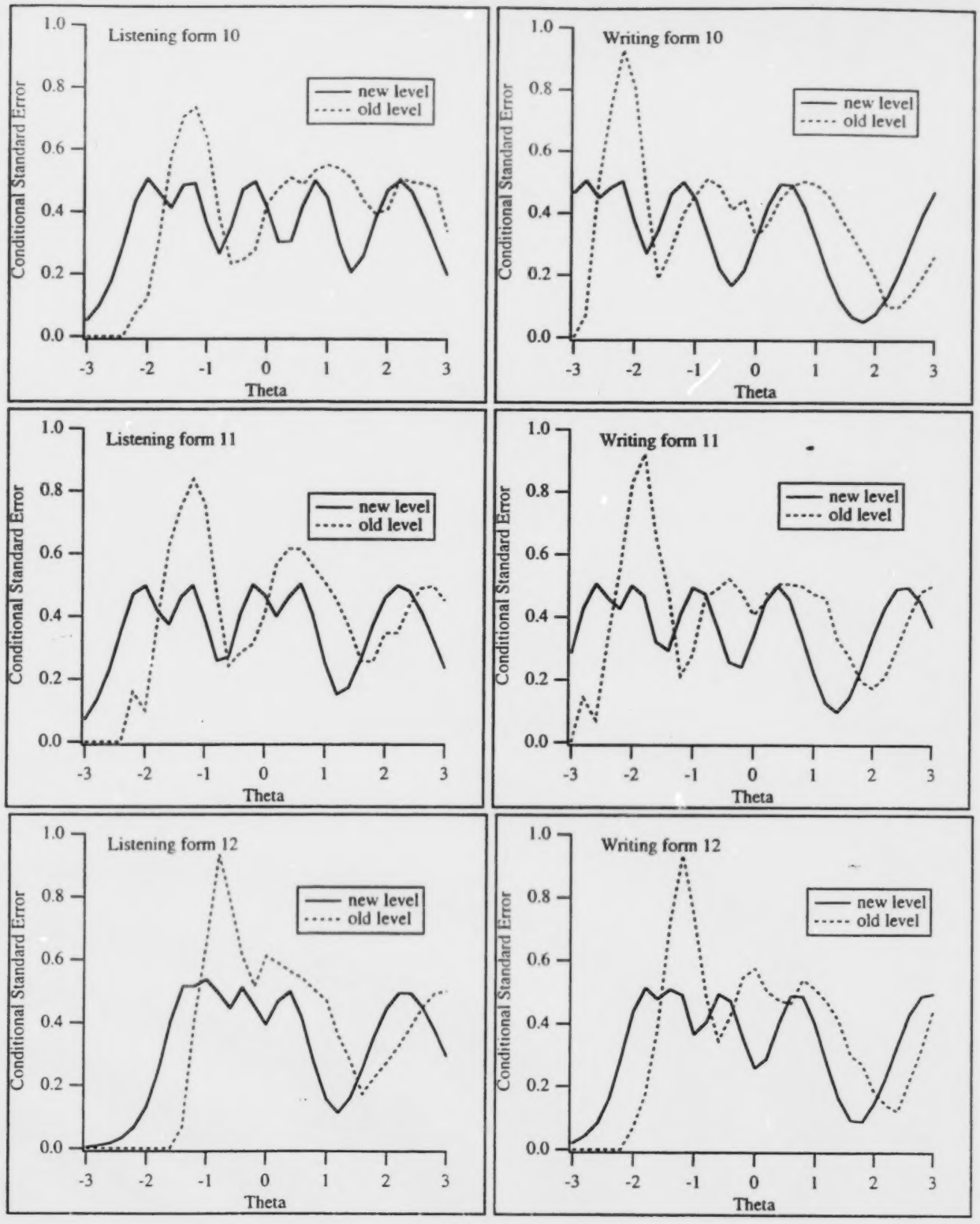


Figure 4. The conditional standard errors for old and new levels on the theta scale.

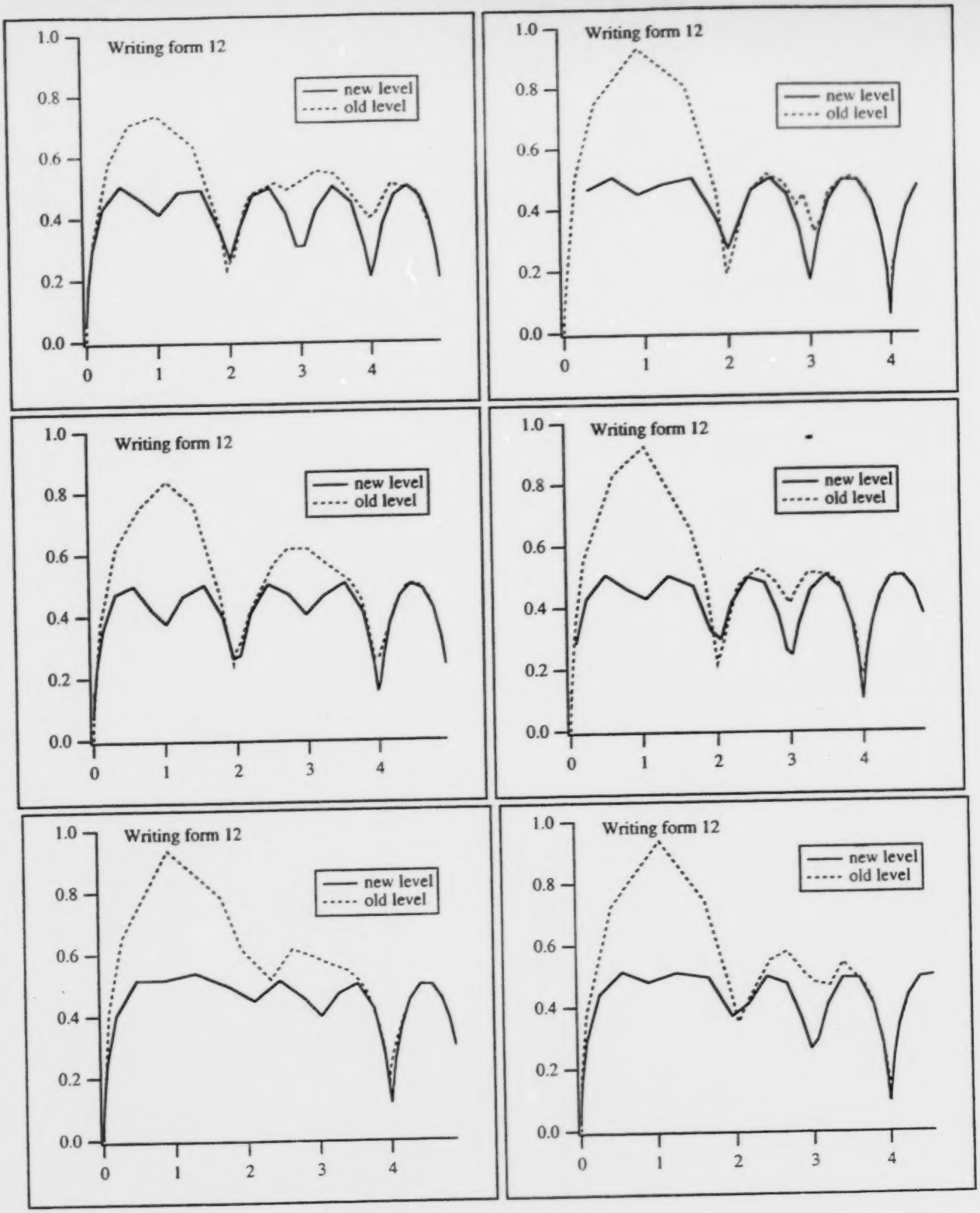


Figure 5. The conditional standard errors for old and new levels on the level score scale.

Table 1. FACETS based marginal distribution for the Listening test.

		Marginal Distribution			
		Old		New	
Form	Level	Estimated	Observed	Estimated	Observed
10	0	0.0257	0.028	0.0052	0.003
	1	0.0363	0.036	0.0266	0.019
	2	0.6069	0.628	0.3272	0.246
	3	0.2872	0.297	0.5329	0.655
	4	0.0432	0.011	0.1068	0.075
	5	0.0007	0.000	0.0013	0.001
11	0	0.0462	0.043	0.0109	0.007
	1	0.0321	0.057	0.0407	0.038
	2	0.5384	0.551	0.3153	0.245
	3	0.2891	0.319	0.4506	0.562
	4	0.0938	0.030	0.1810	0.147
	5	0.0005	0.000	0.0016	0.001
12	0	0.0462	0.094	0.0109	0.008
	1	0.0321	0.042	0.0407	0.037
	2	0.5384	0.371	0.3153	0.164
	3	0.2891	0.428	0.4506	0.553
	4	0.0938	0.065	0.1810	0.238
	5	0.0005	0.000	0.0016	0.001

Table 2. FACETS based marginal distribution for the Writing test.

		Marginal Distribution			
		Old		New	
Form	Level	Estimated	Observed	Estimated	Observed
10	0	0.0343	0.030	0.0054	0.003
	1	0.0098	0.014	0.0245	0.018
	2	0.2450	0.245	0.1730	0.161
	3	0.4295	0.538	0.4249	0.458
	4	0.2808	0.172	0.3675	0.353
	5	0.0006	0.001	0.0047	0.007
11	0	0.0443	0.039	0.0069	0.001
	1	0.0143	0.025	0.0362	0.034
	2	0.2921	0.266	0.1930	0.179
	3	0.3742	0.495	0.3951	0.422
	4	0.2686	0.170	0.3501	0.344
	5	0.0065	0.006	0.0188	0.020
12	0	0.0912	0.091	0.0103	0.005
	1	0.0193	0.027	0.0617	0.049
	2	0.3561	0.328	0.2513	0.240
	3	0.3439	0.461	0.4319	0.467
	4	0.1886	0.093	0.2417	0.235
	5	0.0008	0.001	0.0031	0.004

Table 3. Classification consistency and reliability indices for Listening test.

	Old			New		
Form	p0	pc	kappa	p0	pc	kappa
10	0.6519	0.4547	0.3617	0.6834	0.4032	0.4695
11	0.6138	0.3854	0.3717	0.6508	0.3369	0.4733
12	0.5436	0.3051	0.3433	0.6344	0.3348	0.4504
	var(T)	var(E)	reliability	var(T)	var(E)	reliability
10	0.3130	0.2106	0.5978	0.3494	0.1608	0.6848
11	0.5007	0.2611	0.6573	0.5217	0.1818	0.7416
12	0.7608	0.3306	0.6970	0.5773	0.1948	0.7477

Table 4. Classification consistency and reliability indices for Writing test.

Form	Old			New		
	p0	pc	kappa	p0	pc	kappa
10	0.6945	0.3247	0.5476	0.7638	0.3461	0.6387
11	0.6482	0.2997	0.4980	0.7285	0.3176	0.6021
12	0.6196	0.2894	0.4648	0.6959	0.3120	0.5579
	var(T)	var(E)	reliability	var(T)	var(E)	reliability
10	0.6793	0.1902	0.7812	0.5677	0.1190	0.8267
11	0.7897	0.2244	0.7787	0.6771	0.1372	0.8315
12	0.9226	0.2963	0.7569	0.6635	0.1566	0.8090